Milestone 1: Index Construction Report

This report presents the analytics for the inverted index constructed from the web page corpus. The index was built using the following specifications:

• Tokens: All alphanumeric sequences
• Stop words: Not used (all words indexed)
• Stemming: Porter stemming algorithm
• Important words: Words in bold, headings (h1, h2, h3), and titles are marked as important
• Term frequency: Calculated for each token in each document

Index Analytics

Metric	Value
Number of Indexed Documents	1,212
Number of Unique Tokens	13,126
Total Size of Index on Disk	22029.18 KB

Implementation Details

Index Structure:
The inverted index is stored as a dictionary where each token maps to a dictionary of document IDs. Each posting contains the term frequency (tf) and a flag indicating if the token appears in important contexts (bold, headings, or title).

File Organization:
The index is saved in JSON format with two files:
• inverted_index.json: Contains the main inverted index structure
• doc_mapping.json: Contains mappings between URLs and document IDs

Processing Pipeline:
1. HTML content is parsed using BeautifulSoup to extract text and identify important elements
2. Text is tokenized into alphanumeric sequences
3. Tokens are stemmed using the Porter stemming algorithm
4. Term frequencies are calculated and stored in the inverted index